Committee Machines

授課教師: 張傳育 博士 (Chuan-Yu Chang Ph.D.)
E-mail: chuanyu@yuntech.edu.tw
Tel: (05)5342601 ext. 4516
Office: EB212
Web: http://MIPL.yuntech.edu.tw

Introduction

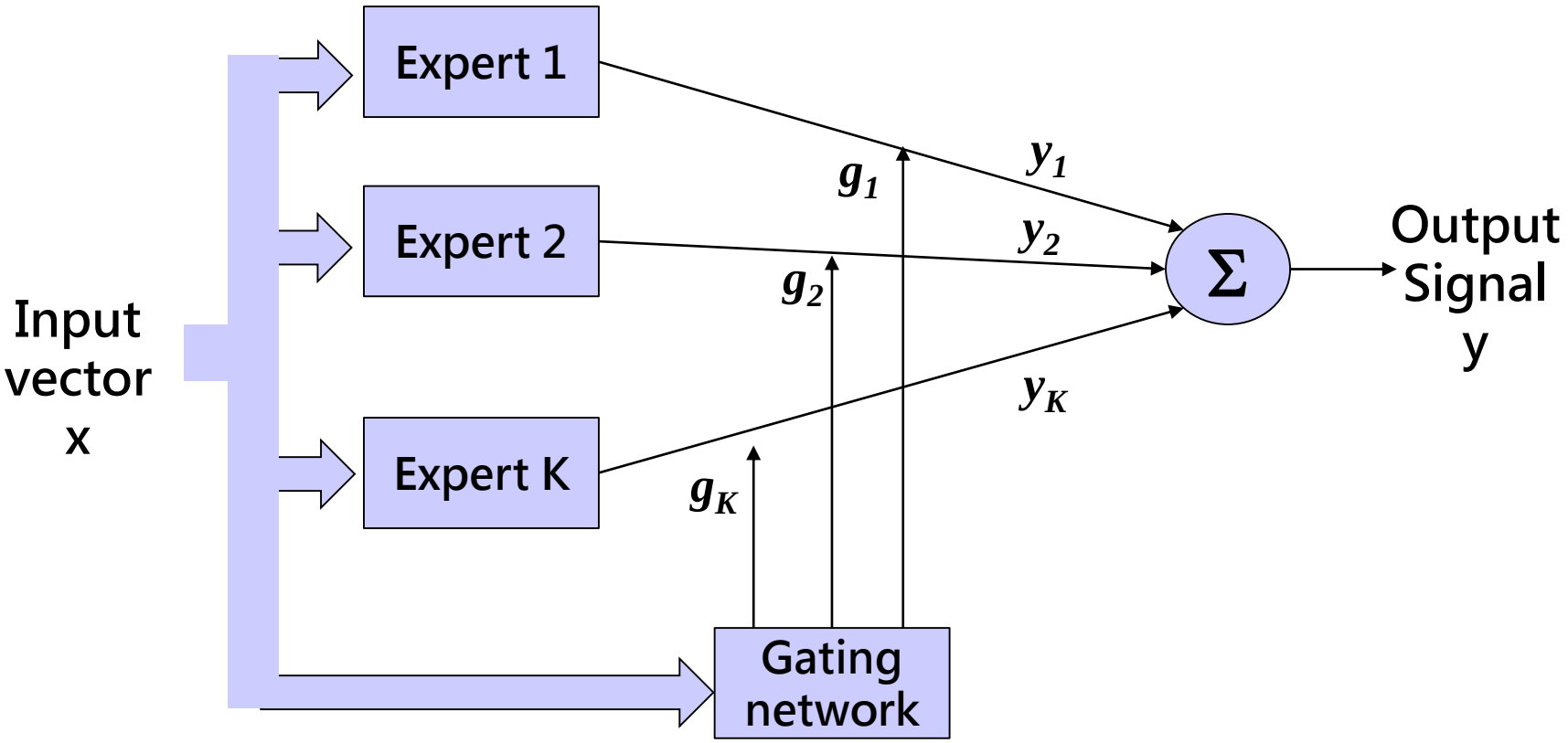
- Principle of divide and conquer
  - A complex computational task is solved by dividing it into a number of computationally simple tasks and then combining the solution to those tasks.
  - In supervised learning, computational simplicity is achieved by distributing the learning task among a number of *experts*.
  - The combination of experts is said to constitute a *committee machine*.
- Committee machine are universal approximators.

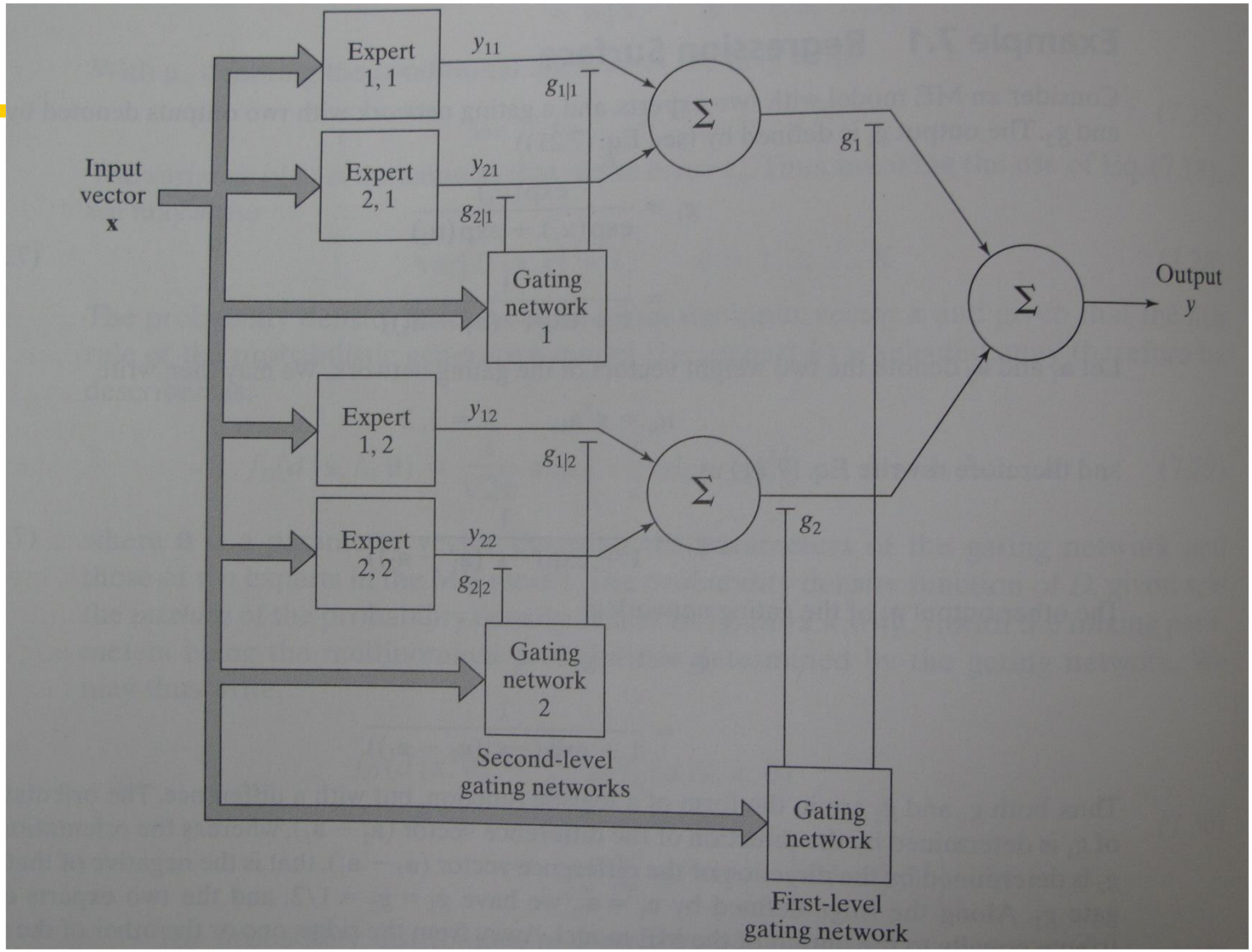
Introduction (cont.)

- They may be classified into two major categories:
  - Static structures
    - The responses of several experts are combined by means of a mechanism that *does not* involve the input signal.
      - Ensemble averaging
        - The outputs of different experts are linearly combined to produce an overall output.
      - Boosting
        - A weak learning algorithm is converted into one that achieves arbitrarily high accuracy.

Introduction (cont.)

- Dynamic structures
  - The input signal is directly involved in actuating the mechanism that integrates the outputs of the individual experts into an overall output.
    - Mixture of experts
      - The individual responses of the experts are nonlinearly combined by means of a single gating network.
    - Hierarchical mixture of experts
      - The individual response of the experts are nonlinearly combined be means of several gating networks arranged in a hierarchical fashion.


Input vector x
Expert 1
Expert 2
Expert K
Gating network
g_1
g_2
g_K
y_1
y_2
y_K
Σ
Output Signal y

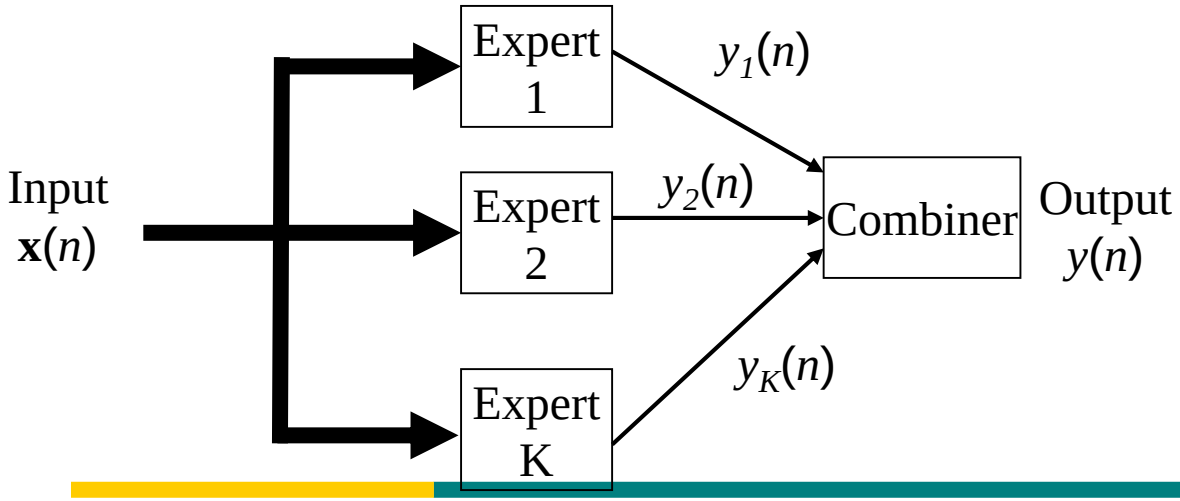

Input vector x
Expert 1, 1
Expert 2, 1
Expert 1, 2
Expert 2, 2
y_{11}
y_{21}
y_{12}
y_{22}
$g_{1|1}$
$g_{2|1}$
$g_{1|2}$
$g_{2|2}$
Gating network 1
Gating network 2
Second-level gating networks
g_1
g_2
Gating network
First-level gating network
Σ
Σ
Σ
Output y

Introduction (cont.)

- The mixture of experts and hierarchical mixture of experts may be viewed as *modular networks*.
  - Modularity
    - If the computation performed by the network can be decomposed into two or more modules that operate on distinct inputs without communicating with each other.(每個**module**獨立工作)
    - The outputs of the modules are mediated by an integrating unit that is not permitted to feed information back to the modules.(再由**integrating unit** 統合結果)
    - The integrating unit
      - Decides how the outputs of the modules should be combined to form the final output of the system
      - Decides which modules should learn which training pattern.

Ensemble Averaging

- Fig. 1 shares a common input and whose individual outputs are somehow combined to produce an overall output y.
- The expectation is that the differently trained experts converge to different local minima on the error surface, and overall performance is improved by combining the outputs in some way.



Boosting (cont.)

- In a committee machine, all the experts in the machine are trained on the same data set, they may differ from each other in the choice of initial conditions used in network training.
- In a boosting machine, the experts are trained on data sets with entirely different distributions.

Boosting (cont.)

- Boosting can be implemented in three ways:
  - Boosting by filtering:
    - Filtering the training examples by different version of weak learning algorithm
      - 優點是與其他方法相比，所需的記憶體較少。
      - 假設在有眾多的data時。
  - Boosting by subsampling:
    - fixed-size training sample
    - 但resample這些data with a probability distribution
  - boosting by reweighting:
    - fixed-size training sample
    - 假設sample為weighted
    - Weak learning根據weighted examples

Boosting (cont.)

- Boosting by Filtering
  - The original idea of boosting
    - Rooted in a distribution free or probably approximately correct (PAC) model of learning.
    - In PAC learning, a learning machine tries to identify an unknown binary concept on the basis of randomly chosen examples of the concept.
    - The goal of the learning machine is to find a hypothesis or prediction rule with an error rate of at most ε, and this should hold uniformly for all input distribution.
    - PAC therefore referred to as a *strong learning model*.
    - 但由於為unknown concept，因此learning有其limitation，因此requirement 為只須1-δ比率可達此要求即可。

Boosting (cont.)

- Weak learning model
  - The learning machine is required to find a hypothesis with an error rate only slightly less than 1/2.(只要求任一者之error rate略小於1/2即可)
  - When a hypothesis guesses a binary label in an entirely random manner on every example, it can be right or wrong with equal probability.
  - It achieves an error rate of exactly 1/2。
- Converting a weak learning model into a strong learning model, by modifying the distribution of examples.

Boosting (cont.)

- In boosting by filtering, the committee machine consists of three experts (subhypotheses)
  - The algorithm used to train them is called “boosting algorithm”.
  - The three experts are arbitrarily labeled 1st, 2nd, and 3rd.
  - The three experts are individually trained as follows:
    - 1. The first expert is trained on N_1 examples (用掉N_1 data)
    - 2. The trained first expert作為filter for selecting pattern for training expert 2
      - random choose coin face (head vs tail)
      - a)若是正面(head)，則忽略或丟棄可被第一個expert正確分類的樣本，將被錯誤分類的樣本加入第二個 expert的training set。
      - b)若是反面(tail):類似(a) ，但選可被正確classified pattern作為第二個 expert的training example。

用掉N_2 data, 選出N_1個

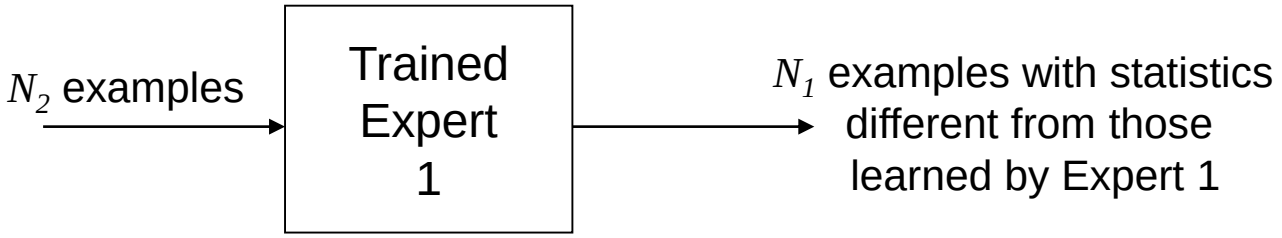
    - Repeat直到N_1個pattern被選出,train expert 2

表示expert 1 之data 與expert 2 完全無關(orthogonal)

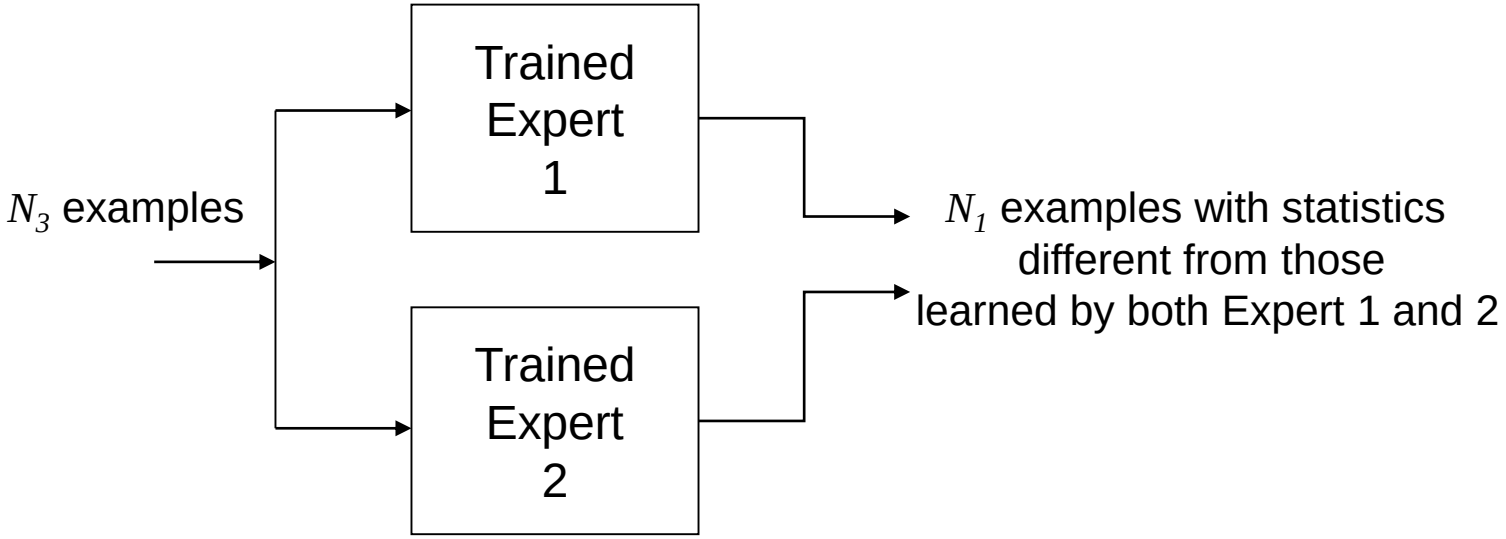
Boosting (cont.)

- 當expert 1 及expert 2 均trained 好。用其過濾 expert 3 所須之training data
  - pass a pattern to 1及2，若它們的classification一致，則discard此pattern，若不一致，留此pattern以train expert 3。
  - continue，直到N_1個pattern被選出。
  - 特性：
    - (a) committee machine requires a large set of examples for its operation，但只小部份真正用 training。
    - (b) 由於用先前之experts來濾出須training之資料， focus on “hard-to-learn” parts。

Boosting (cont.)



Filtering of examples performed by Expert 1



Filtering of examples performed by Expert 1 and 2

Boosting (cont.)

- AdaBoost
  - A practical limitation of boosting by filtering is that it often requires a large training sample.
  - 可使用AdaBoost來解決上述問題。
    - 屬於boosting by resampling
    - It permits the training data to be reused.
    - The goal is to find a final mapping function or hypothesis with low error rate relative to a given distribution $\mathcal{D}$ over the labeled training examples.

Boosting (cont.)

- AdaBoost differs from other boosting algorithms in two respects:
  - AdaBoost adjusts *adaptively* to the errors of the weak hypothesis returned by the weak learning model.
  - The bound on performance of AdaBoost depends only on the performance of the weak learning model on those distributions that are actually generated during the learning process.

Boosting (cont.)

- Summary of AdaBoost
  - Input: Training sample $\{(\mathbf{x}_i, d_i)\}^N_{i=1}$
    Distribution $\mathscr{D}$ over the N labeled examples.
    Weak learning model
    Integer T specifying the number of iterations of the algorithm
  - Initialization: Do the following for n=1, 2,…, T:
    - 1. Call the weak learning model, providing it with the distribution $\mathscr{D}_n$.
    - 2. Get back hypothesis $\mathscr{F}_n$: $\mathbf{X} \rightarrow Y$
    - 3. Calculate the error of hypothesis $\mathscr{F}_n$:

$$\varepsilon_n = \sum_{i:\mathscr{F}_n(\mathbf{x}_i)\neq d_i} \mathscr{D}_n(i)$$

    - Set $\beta_n = \dfrac{\varepsilon_n}{1-\varepsilon_n}$
    - Update the distribution $\mathscr{D}_n$:

$$\mathscr{D}_{n+1}(i) = \frac{\mathscr{D}_n(i)}{Z_n} \times \begin{cases} \beta_n & \text{if } \mathscr{F}_n(\mathbf{x}_i) = d_i \\ 1 & \text{otherwise} \end{cases}$$

  - Output: The final hypothesis is

$$\mathscr{F}_n(\mathbf{x}) = \arg\max_{d\in\mathscr{D}} \sum_{n:\mathscr{F}_n(\mathbf{x})=d} \log\frac{1}{\beta_n}$$